

PCT

WELTORGANISATION FÜR GEISTIGES EIGENTUM
Internationales Büro



INTERNATIONALE ANMELDUNG VERÖFFENTLICHT NACH DEM VERTRAG ÜBER DIE
INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT)

| | | |
|---|-----------|---|
| (51) Internationale Patentklassifikation ⁶ : G10L 5/06 | A1 | (11) Internationale Veröffentlichungsnummer: WO 99/14740 (43) Internationales Veröffentlichungsdatum: 25. März 1999 (25.03.99) |
|---|-----------|---|

(21) Internationales Aktenzeichen: PCT/DE98/02632

(22) Internationales Anmeldedatum: 7. September 1998 (07.09.98)

(30) Prioritätsdaten:
197 40 911.3 17. September 1997 (17.09.97) DE

(71) Anmelder (für alle Bestimmungsstaaten ausser US): SIEMENS
AKTIENGESELLSCHAFT [DE/DE]; Wittelsbacherplatz 2,
D-80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder (nur für US): WITSCHER, Petra [DE/DE];
Helmpertstrasse 15/1, D-80687 München (DE). HÖGE,
Harald [DE/DE]; Obertaxetweg 6b, D-82131 Gauting (DE).

(74) Gemeinsamer Vertreter: SIEMENS AKTIENGE-
SELLSCHAFT; Postfach 22 16 34, D-80506 München
(DE).

(81) Bestimmungsstaaten: CN, JP, US, europäisches Patent (AT,
BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU,
MC, NL, PT, SE).

Veröffentlicht

*Mit internationalem Recherchenbericht.
Vor Ablauf der für Änderungen der Ansprüche zugelassenen
Frist; Veröffentlichung wird wiederholt falls Änderungen
eintreffen.*

(54) Title: METHOD FOR DETERMINING THE PROBABILITY OF THE OCCURRENCE OF A SEQUENCE OF AT LEAST TWO
WORDS IN A SPEECH RECOGNITION PROCESS

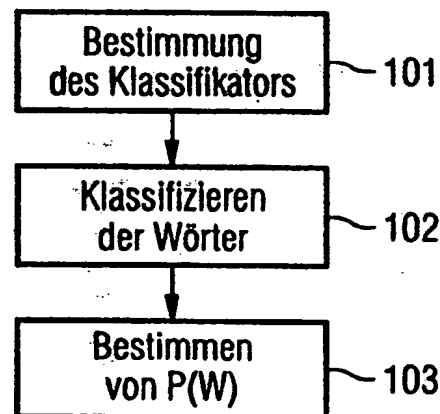
(54) Bezeichnung: VERFAHREN ZUR BESTIMMUNG EINER WAHRSCHEINLICHKEIT FÜR DAS AUFTRETEN EINER FOLGE
VON MINDESTENS ZWEI WÖRTERN BEI EINER SPRACHERKENNUNG

(57) **Abstract**

Statistical speech models are used to improve the recognition rate of automatic speech recognition systems. The inventive method makes it possible to construct speech models with small volumes of text or without a text. Linguistic classes are formed and the speech model, i.e. an a-priori probability of word sequences, is determined based on the grammatical structure of the text and depending on the vocabulary of the text. Set probabilities for words and for linguistic classes are determined for this purpose. In addition to creating a new speech model for a predetermined area of application, the inventive method also provides a means of adapting a pre-existing speech model to a new area of application.

(57) **Zusammenfassung**

Um die Erkennungsrate von automatischen Spracherkennungssystemen zu verbessern, werden statistische Sprachmodelle verwendet. Mit dem angegebenen Verfahren ist es möglich, Sprachmodelle auf kleinen Textmengen bzw. ganz ohne einen Text aufzubauen. Es werden linguistische Klassen gebildet und das Sprachmodell, also eine a-priori Wahrscheinlichkeit für Wortfolgen, basierend auf der grammatikalischen Struktur des Textes und abhängig vom Vokabular des Textes ermittelt. Dazu werden bedingte Wahrscheinlichkeiten für Wörter und für linguistische Klassen bestimmt. Neben der Neuerschaffung eines Sprachmodells für einen vorgegebenen Anwendungsbereich enthält das Verfahren ebenso eine Möglichkeit zur Anpassung eines bereits bestehenden Sprachmodells an einen neuen Anwendungsbereich.



101...DETERMINATION OF CLASSIFIER
102...WORD CLASSIFICATION
103...DETERMINATION OF P (W)

LEDIGLICH ZUR INFORMATION

Codes zur Identifizierung von PCT-Vertragsstaaten auf den Kopfbögen der Schriften, die internationale Anmeldungen gemäss dem PCT veröffentlichen.

| | | | | | | | |
|----|------------------------------|----|-----------------------------|----|-----------------------------|----|------------------------|
| AL | Albanien | ES | Spanien | LS | Lesotho | SI | Slowenien |
| AM | Armenien | FI | Finnland | LT | Litauen | SK | Slowakei |
| AT | Österreich | FR | Frankreich | LU | Luxemburg | SN | Senegal |
| AU | Australien | GA | Gabun | LV | Lettland | SZ | Swasiland |
| AZ | Aserbaidschan | GB | Vereinigtes Königreich | MC | Monaco | TD | Tschad |
| BA | Bosnien-Herzegowina | GE | Georgien | MD | Republik Moldau | TG | Togo |
| BB | Barbados | GH | Ghana | MG | Madagaskar | TJ | Tadschikistan |
| BE | Belgien | GN | Guinea | MK | Die ehemalige jugoslawische | TM | Turkmenistan |
| BF | Burkina Faso | GR | Griechenland | | Republik Mazedonien | TR | Türkei |
| BG | Bulgarien | HU | Ungarn | ML | Mali | TT | Trinidad und Tobago |
| BJ | Benin | IE | Irland | MN | Mongolei | UA | Ukraine |
| BR | Brasilien | IL | Israel | MR | Mauretanien | UG | Uganda |
| BY | Belarus | IS | Island | MW | Malawi | US | Vereinigte Staaten von |
| CA | Kanada | IT | Italien | MX | Mexiko | | Amerika |
| CF | Zentralafrikanische Republik | JP | Japan | NE | Niger | UZ | Usbekistan |
| CG | Kongo | KE | Kenia | NL | Niederlande | VN | Vietnam |
| CH | Schweiz | KG | Kirgisistan | NO | Norwegen | YU | Jugoslawien |
| CI | Côte d'Ivoire | KP | Demokratische Volksrepublik | NZ | Neuseeland | ZW | Zimbabwe |
| CM | Kamerun | | Korea | PL | Polen | | |
| CN | China | KR | Republik Korea | PT | Portugal | | |
| CU | Kuba | KZ | Kasachstan | RO | Rumänien | | |
| CZ | Tschechische Republik | LC | St. Lucia | RU | Russische Föderation | | |
| DE | Deutschland | LI | Liechtenstein | SD | Sudan | | |
| DK | Dänemark | LK | Sri Lanka | SE | Schweden | | |
| EE | Estland | LR | Liberia | SG | Singapur | | |

Beschreibung

VERFAHREN ZUR BESTIMMUNG EINER WAHRSCHEINLICHKEIT FÜR DAS AUFTRETEN EINER FOLGE VON MINDESTENS ZWEI WÖRTERN BEI EINER SPRACHERKENNUNG

5

Die Erfindung betrifft ein Verfahren zur Bestimmung einer Wahrscheinlichkeit für das Auftreten einer Folge von mindestens zwei Wörtern bei einer Spracherkennung durch einen
10 Rechner.

Ein Verfahren zur Spracherkennung ist aus [1] bekannt. Dabei ist es als ein Bestandteil der Worterkennung üblich, die Brauchbarkeit einer Folge aus mindestens einem Wort
15 anzugeben. Ein Maß für diese Brauchbarkeit ist eine Wahrscheinlichkeit.

Ein statistisches Sprachmodell ist aus [2] bekannt. So kennzeichnet die Wahrscheinlichkeit $P(W)$ für eine Wortfolge W
20 im Rahmen der Spracherkennung, vorzugsweise großer Vokalmengen, allgemein ein (statistisches) Sprachmodell. Die Wahrscheinlichkeit $P(W)$ (sog.: Wortfolgewahrscheinlichkeit) wird angenähert durch ein N-Gramm-Sprachmodell $P_N(W)$:

$$25 \quad P_N(W) = \prod_{i=0}^n P(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-N+1}), \quad (0-1)$$

wobei

w_i das i -te Wort der Folge W mit ($i=1..n$),

n die Anzahl der Wörter w_i der Folge W

bezeichnen.

30 Für $N=2$ ergeben sich aus Gleichung (0-1) sogenannte Bigramme.

Ferner ist bekannt, bei der Spracherkennung, vorzugsweise im kommerziellen Umfeld, ein im Vokabular beschränktes Anwendungsgebiet (Domäne) zu verwenden. Texte verschiedener
35 Domänen unterscheiden sich nicht nur im Hinblick auf ihr

jeweiliges Vokabular, sondern auch hinsichtlich ihres jeweiligen Satzaufbaus. Für ein Trainieren eines Sprachmodells für eine spezielle Domäne ist eine entsprechend große Menge an Texten (Textmaterial, Textkorpus) notwendig, das jedoch in der Praxis nur selten vorhanden oder nur mit

5 immensem Aufwand zu gewinnen ist.

Ein linguistisches Lexikon ist aus [4] bekannt. Dabei handelt es sich um eine auf einem Rechner verfügbare Zusammenstellung

10 möglichst vieler Wörter einer Sprache zum Zwecke des Nachschlagens von linguistischen Eigenschaften anhand eines Suchprogramms. Für jeden Worteintrag (sog. Wortvollform) können die für diese Wortvollform relevanten linguistischen Merkmale und die zutreffenden Belegungen, also die

15 linguistischen Werte, entnommen werden.

Aus [3] ist bekannt, linguistische Klassen zu verwenden. Wörter eines Satzes können auf unterschiedliche Arten in linguistischen Merkmalen und linguistischen Werten zugeordnet

20 werden. In Tabelle 1 sind beispielhaft verschiedene linguistischen Merkmale und die zugehörigen Werte dargestellt (weitere Beispiele sind in [3] angegeben).

| ling. Merkmal | linguistische Werte |
|---------------------|--|
| Kategorie | Substantiv, Verb, Adjektiv, Artikel, Pronomen, Adverb, Konjunktion, Präposition, etc. |
| Typ des Substantivs | abstrakt, tierisch, als Teil des Körpers, konkret, menschlich, örtlich, materiell, als Maß, pflanzlich, zeitlich, etc. |
| Art des Pronomens | demonstrativ, indefinit, interrogativ, possessiv, etc. |

Tabelle 1: Beispiele für ling. Merkmale und ling. Werte

Basierend auf linguistischen Merkmalen

$$(f_1, \dots, f_m) \quad (0-2)$$

5 und linguistischen Werten

$$(v_{11} \dots v_{1j}) \dots (v_{m1} \dots v_{mj}) \quad (0-3)$$

10 wird jedem Wort mindestens eine linguistische Klasse
zugewiesen, wobei folgende Abbildungsvorschrift F Anwendung
findet:

$$(C_1, \dots, C_k) = F \left((f_1, v_{11}, \dots, v_{1j}) \dots (f_m, v_{m1}, \dots, v_{mj}) \right) \quad (0-4)$$

15 wobei

f_m ein linguistisches Merkmal,
 m die Anzahl linguistischer Merkmale,
 $v_{m1} \dots v_{mj}$ die linguistischen Werte des linguistischen
Merkmals f_m ,

20 j die Anzahl der linguistischen Werte,
 C_i die linguistische Klasse mit $i=1..k$,
 k die Anzahl der linguistischen Klassen,
 F eine Abbildungsvorschrift (Klassifikator)
von linguistischen Merkmalen und

25 linguistischen Werten auf linguistische
Klassen

bezeichnen.

30 Zur Veranschaulichung der linguistischen Klasse, des
linguistischen Merkmals (engl. feature), des linguistischen
Werts (engl. value) und der Klassen-Bigramm-
Wahrscheinlichkeit wird nachfolgend ein Beispiel erläutert.

Ausgegangen wird von dem deutschen Satz:

"der Bundestag setzt seine Debatte fort"

Der Artikel "der" (also das erste Wort) kann im Deutschen in sechs linguistische Klassen (fortan nur: Klassen) unterteilt werden, wobei sich die Klassen in Numerus, Genus und Kasus unterscheiden. Die folgende Tabelle 2 veranschaulicht diesen Zusammenhang:

| C_i | Kategorie | Numerus | Genus | Kasus |
|-------|-----------|----------|------------|-----------|
| C_1 | Artikel | Singular | maskulinum | Nominativ |
| C_2 | Artikel | Singular | femininum | Genitiv |
| C_3 | Artikel | Singular | femininum | Dativ |
| C_4 | Artikel | Plural | femininum | Genitiv |
| C_5 | Artikel | Plural | maskulinum | Genitiv |
| C_6 | Artikel | Plural | neutrum | Genitiv |

Tabelle 2: Klassen C_i für das Wort "der"

10

Für das deutsche Substantiv "Bundestag" (zweites Wort des obigen Beispielsatzes) folgt analog Tabelle 3:

| C_i | Kategorie | Numerus | Genus | Kasus |
|-------|------------|----------|------------|-----------|
| C_7 | Substantiv | Singular | maskulinum | Nominativ |
| C_8 | Substantiv | Singular | maskulinum | Akkusativ |
| C_9 | Substantiv | Singular | maskulinum | Dativ |

Tabelle 3: Klassen C_i für das Wort "Bundestag"

15

In diesem Beispiel folgt nun im Hinblick auf Klassen-Bigramme, also Bigrammen angewandt auf linguistische Klassen, daß die Klasse C_1 , gefolgt von der Klasse C_7 die richtige Kombination von Kategorie, Numerus, Kasus und Genus bezüglich des Beispielsatzes darstellt. Wenn Häufigkeiten real

20

vorkommender Klassen-Bigramme aus vorgegebenen Texten bestimmt werden, so folgt für das obige Klassen-Bigramm C_1-C_7 ein zahlreiches Auftreten, da diese Kombination in der deutschen Sprache oft vorkommt, wohingegen anderer Klassen-Bigramme, z.B. die Kombination C_2-C_8 in der deutschen Sprache wegen unterschiedlicher Geni nicht erlaubt ist. Die sich aus den auf diese Art gefundenen Häufigkeiten ergebenden Klassen-Bigramm-Wahrscheinlichkeiten sind entsprechend hoch (bei oftmaligem Auftreten) bzw. niedrig (falls nicht zulässig).

10

Die Aufgabe der Erfindung besteht darin, ein Verfahren zur Bestimmung einer Wahrscheinlichkeit für das Auftreten einer Folge von mindestens einem Wort bei einer Spracherkennung anzugeben, wobei ein Sprachmodell für eine Domäne aus wenig Textmaterial bestimmt wird.

15

Hierbei sei angemerkt, daß eine Bestimmung des Sprachmodells ausdrücklich die Erzeugung als auch eine Adaption eines bestehenden Sprachmodells umfaßt.

20

Die Aufgabe wird gemäß den Merkmalen des Patentanspruchs 1 gelöst.

Erfindungsgemäß angegeben wird ein Verfahren zur Bestimmung einer Wahrscheinlichkeit für das Auftreten einer Folge von mindestens zwei Wörtern bei einer Spracherkennung durch einen Rechner. Eine Sprache weist linguistische Klassen

30

$$(C_1, \dots, C_k) \quad (1)$$

gemäß

$$(C_1, \dots, C_k) = F \left((f_1, v_{11}, \dots, v_{1j}) \dots (f_m, v_{m1}, \dots, v_{mj}) \right) \quad (2),$$

35

wobei

6

f_m ein linguistisches Merkmal,
 m die Anzahl linguistischer Merkmale,
 $v_{m1} \dots v_{mj}$ die linguistischen Werte des linguistischen Merkmals f_m ,

5 j die Anzahl der linguistischen Werte,
 C_i die linguistische Klasse mit $i=1..k$,
 k die Anzahl der linguistischen Klassen,
 F eine Abbildungsvorschrift (Klassifikator)
10 von linguistischen Merkmalen und linguistischen Werten auf linguistische Klassen

bezeichnen, auf.

Dabei wird einem Wort mindestens eine der linguistischen
15 Klassen zugeordnet. Eine Wahrscheinlichkeit $P(W)$ für das Auftreten der Folge von mindestens zwei Wörtern ergibt sich unter Verwendung von Bigrammen zu

$$P(W) \approx \prod_{i=1}^n \sum_{C_i} \sum_{C_{i-1}} P(w_i | C_i) \times P(C_i | C_{i-1}) \times P(C_{i-1} | w_{i-1}) \quad (3)$$

20

wobei

W die Folge von mindestens zwei Wörtern,
 w_i das i -te Wort der Folge W mit $(i=1..n)$,
 n die Anzahl der Wörter w_i der Folge W ,
25 C_i eine linguistische Klasse, die zu einem Wort w_i gehört,
 C_{i-1} eine linguistische Klasse, die zu einem Wort w_{i-1} gehört,
 \sum_{C_i} die Summe über alle linguistischen Klassen
30 C , die zu einem Wort w_i gehören,
 $P(w_i | C_i)$ die bedingte Wortwahrscheinlichkeit,
 $P(C_i | C_{i-1})$ die Wahrscheinlichkeit für Bigramme (auch: Klassen-Bigramme-Wahrscheinlichkeit),

$P(C_{i-1}|w_{i-1})$ die bedingte Klassenwahrscheinlichkeit bezeichnen.

- Hierbei sei angemerkt, daß sich der Term C_i auf eine der
- 5 mindestens einen linguistischen Klasse bezieht, die dem Wort w_i aus der Wortfolge W zugeordnet wird. Das gleiche gilt entsprechend für den Term C_{i-1} . Beispielsweise handelt es sich bei der Klassen-Bigramm-Wahrscheinlichkeit um die Wahrscheinlichkeit dafür, daß das Wort w_i einer ersten
- 10 linguistischen Klasse angehört unter der Bedingung, daß das vorhergehende Wort w_{i-1} einer zweiten linguistischen Klasse angehört (siehe hierzu einleitendes Beispiel und Erläuterung hierzu).
- 15 Die Wahrscheinlichkeiten $P(w_i|C_i)$ und $P(C_i|C_{i-1})$, die eingesetzt in Gleichung (3) ein sogenanntes Basissprachmodell ergeben, können aus einem Textkorpus, also aus einem vorgegebenem Text mit vorgegebenem Umfang, bestimmt werden.
- 20 Sprachmodelle, die auf linguistischen Klassen basieren, bieten insbesondere für eine Adaption entscheidende Vorteile. Das hier vorgestellte Verfahren verwendet die in den Sprachmodellen enthaltenen linguistischen Eigenschaften.
- 25 Eine Weiterbildung besteht darin, daß für einen neuen Text anhand eines vorbestimmten Basissprachmodells die Wahrscheinlichkeit $P(C_i|C_{i-1})$ in das Basissprachmodell für den neuen Text übernommen wird.
- 30 Wahrscheinlichkeiten für Klassen-Bigramme des Basissprachmodells (siehe [3] und Erläuterung in der Einleitung) stellen einerseits eine grammatikalische Struktur des Trainingstextes dar und sind andererseits vom Vokabular unabhängig. Geht man davon aus, daß die neue Domäne von
- 35 ähnlicher Textstruktur (grammatikalischer Struktur) wie der ursprüngliche Trainingstext für das Basissprachmodell ist, so

ist es zweckmäßig, die Wahrscheinlichkeit für die Klassen-Bigramme $P(C_i|C_{i-1})$ aus dem Basissprachmodell unverändert zu übernehmen.

- 5 Das Vokabular für die neue Domäne, für die ein Sprachmodell bestimmt wird, wird mit Hilfe eines vorgegebenen linguistischen Lexikons und unter Verwendung eines Klassifikators F gemäß Gleichung (2) bearbeitet. Für jedes neue Wort aus dem Text wird automatisch mindestens eine
- 10 linguistische Klasse bestimmt. Zur detaillierten Beschreibung von linguistischen Klassen, linguistischen Merkmalen und linguistischen Werten siehe [3], zum linguistischen Lexikon siehe [4] und/oder jeweils die Einleitung.
- 15 Eine andere Weiterbildung besteht darin, die Wahrscheinlichkeit $P(w_i|C_i)$ nach mindestens einer der folgenden Möglichkeiten zu bestimmen:
- 20 a) die Wahrscheinlichkeit $P(w_i|C_i)$ wird anhand des Textes bestimmt;
- b) die Wahrscheinlichkeit $P(w_i|C_i)$ wird für ein Wort w_i mit Hilfe einer vorgegebenen Wahrscheinlichkeit $P(w_i)$ bestimmt;
- 25 c) die Wahrscheinlichkeit $P(w_i|C_i)$ wird unter Verwendung einer Wortliste bestimmt.

Eine zusätzliche Weiterbildung besteht darin, anhand der

30 ermittelten Wahrscheinlichkeit $P(w_i|C_i)$ das Basissprachmodell anzupassen. Dies geschieht vorzugsweise derart, daß diese ermittelten Wahrscheinlichkeiten $P(w_i|C_i)$ in das Basissprachmodell aufgenommen werden.

Auch ist es eine Weiterbildung der Erfindung, die Wahrscheinlichkeit $P(C_{i-1}|w_{i-1})$ anhand der Wahrscheinlichkeit $P(w_i|C_i)$ wie folgt zu bestimmen:

$$5 \quad P(C_i|w_i) = K \times P(w_i|C_i) \times P(C_i) \quad (4)$$

wobei

$$K = \left(\sum_{C_i} P(w_i|C_i) \times P(C_i) \right)^{-1} \quad (5)$$

10

einen Normalisierungsfaktor bezeichnet.

Eine andere Weiterbildung der Erfindung besteht darin, eine entsprechende Folge von mindestens einem Wort zu erkennen,
15 falls die Wahrscheinlichkeit $P(W)$ oberhalb einer vorgegebenen Schranke liegt. Ist dies nicht der Fall, so wird eine vorgegebene Aktion durchgeführt. Diese vorgegebene Aktion ist z.B. die Ausgabe einer Fehlermeldung oder der Abbruch des Verfahrens.

20

Im Rahmen einer anderen Weiterbildung bezieht sich der Text auf einen vorgegebenen Anwendungsbereich, eine sogenannte Domäne.

25 Im Rahmen der Erfindung ist es besonders von Vorteil, daß das vorgestellte Verfahren einen neuen Text von nur geringem Umfang für die Bestimmung eines Sprachmodells einer neuen Domäne benötigt.

30 Auch ist von Vorteil, daß Listen von neuen Wörtern (mit oder ohne Angabe der Wahrscheinlichkeit $P(w_i)$) verwendet werden können. Domänenbezogene Spracherkennung spielt in der Praxis eine große Rolle. Das Verfahren begegnet somit einer realen Anforderung und hat sich im Experiment als geeignet und
35 überaus nützlich erwiesen. Greift man auf das

Basissprachmodell zurück, so reduziert sich die Anzahl der neu abzuschätzenden Wahrscheinlichkeiten (Abschätzung nur von $P(w_i|C_i)$ notwendig) erheblich.

- 5 Weiterbildungen der Erfindung ergeben sich auch aus den abhängigen Ansprüchen.

Anhand der folgenden Figuren werden Ausführungsbeispiele der Erfindung näher dargestellt.

10

Es zeigen

- Fig.1 ein Blockdiagramm, das Schritte eines Verfahrens zur Bestimmung einer Wahrscheinlichkeit für das Auftreten einer Folge von mindestens einem Wort bei einer Spracherkennung durch einen Rechner umfaßt,

15

- Fig.2 ein erstes Adaptionsverfahren zur Bestimmung der Wahrscheinlichkeit $P(w_i|C_i)$,

20

- Fig.3 ein zweites Adaptionsverfahren zur Bestimmung der Wahrscheinlichkeit $P(w_i|C_i)$.

- Fig.1 stellt einzelne Schritte des Verfahrens zur Bestimmung einer Wahrscheinlichkeit für das Auftreten einer Folge von mindestens zwei Wörtern bei einer Spracherkennung durch einen Rechner dar. In einem Schritt 101 wird Sprache in linguistische Klassen

25

$$30 \quad (C_1, \dots, C_k) \quad (1)$$

unterteilt:

$$(C_1, \dots, C_k) = F \left((f_1, v_{11}, \dots, v_{1j}) \dots (f_m, v_{m1}, \dots, v_{mj}) \right) \quad (2),$$

35

wobei

- f_m ein linguistisches Merkmal,
 m die Anzahl linguistischer Merkmale,
 $v_{m1} \dots v_{mj}$ die linguistischen Werte des linguistischen Merkmals f_m ,
5 j die Anzahl der linguistischen Werte,
 C_i die linguistische Klasse mit $i=1..k$,
 k die Anzahl der linguistischen Klassen,
 F eine Abbildungsvorschrift (Klassifikator)
10 von linguistischen Merkmalen und linguistischen Werten auf linguistische Klassen

bezeichnen.

15 Eine ausführliche Erklärung der linguistischen Merkmale und der linguistischen Werte findet sich in [3], z.B. auf Seite 1201 in Tabelle 4 wird eine beispielhafte Aufstellung von linguistischen Merkmalen mit zugehörigen linguistischen Werten abhängig von unterschiedlichen Kategorien gezeigt.

20 In einem Schritt 102 wird einem Wort mindestens eine der linguistischen Klassen zugeordnet. Wie in [3] beschrieben ist, kann einem Wort eine oder mehrere der linguistischen Klassen zugeordnet werden.

25 Schließlich wird in einem Schritt 103 die Wahrscheinlichkeit $P(W)$ für das Auftreten der Folge von mindestens einem Wort bestimmt durch

$$P(W) \approx \prod_{i=1}^n \sum_{C_i} \sum_{C_{i-1}} P(w_i | C_i) \times P(C_i | C_{i-1}) \times P(C_{i-1} | w_{i-1}) \quad (3)$$

30

wobei

- W die Folge von mindestens zwei Wörtern,
 w_i das i -te Wort der Folge W mit ($i=1..n$),
 n die Anzahl der Wörter w_i der Folge W ,

12

C_i eine linguistische Klasse, die zu einem Wort w_i gehört,
 C_{i-1} eine linguistische Klasse, die zu einem Wort w_{i-1} gehört,
5 $\sum C_i$ die Summe über alle linguistischen Klassen C , die zu einem Wort w_i gehören,
 $P(w_i|C_i)$ die bedingte Wortwahrscheinlichkeit,
 $P(C_i|C_{i-1})$ die Wahrscheinlichkeit für Bigramme (auch:
10 Klassen-Bigramme, Bigramm-Wahrscheinlichkeit),
 $P(C_{i-1}|w_{i-1})$ die bedingte Klassenwahrscheinlichkeit
bezeichnen.

Die Gleichung (3) besteht aus einer kaskadierten

15 Multiplikation dreier Komponenten

$P(C_i|C_{i-1})$, $P(w_i|C_i)$ und $P(C_{i-1}|w_{i-1})$,

die nachfolgend im einzelnen bestimmt werden.

20

Bestimmung der Wahrscheinlichkeit $P(C_i|C_{i-1})$:

Vokabeln des Textes für die neue Domäne werden mit Hilfe
25 eines linguistischen Lexikons unter Verwendung eines
Klassifikators F , wie in Gleichung (2) gezeigt,
linguistischen Klassen zugeordnet. Dabei wird jedes neue Wort
automatisch mindestens einer linguistischen Klasse
zugeordnet. Ein Basissprachmodell umfaßt Wahrscheinlichkeiten
30 für Klassen-Bigramme [3], wobei diese Wahrscheinlichkeit
einerseits eine grammatikalische Struktur repräsentiert,
andererseits von den einzelnen Wörtern unabhängig ist. Wird
nun davon ausgegangen, daß die Domäne, also der spezielle
anwendungsbezogene Themenbereich, von einer ähnlichen

35 Textstruktur ist wie der dem Basissprachmodell

zugrundeliegende Trainingstext, so wird die Wahrscheinlichkeit für Klassen-Bigramme $P(C_i|C_{i-1})$ aus dem Basissprachmodell unverändert übernommen. Eine solche automatische Zuordnung ist möglich, da, weitgehend unabhängig vom Inhalt eines Textes, davon ausgegangen werden kann, daß linguistische Merkmale und linguistische Werte, also eine Abbildung in linguistische Klassen (siehe Gleichung (2)), eigentümlich für eine Sprache sind und daher vorteilhaft von einem Themenbereich (Domäne) auf einen anderen Themenbereich übertragen werden können. Grammatikalische Strukturen vieler Trainingstexte sind ähnlich bzw. gleich, unabhängig von dem den jeweiligen Trainingstexten zugrundeliegenden Themenbereich (Domäne). Die Wahrscheinlichkeit $P(C_i|C_{i-1})$, die für das Basissprachmodell aufgrund von vorausgegangenen Trainingsdurchgängen bestimmt wurde, stellt in Bezug auf die Klassenzuordnung des Vokabulars des neuen Textes ein Grundwissen dar, das nicht erneut bestimmt werden muß, sondern, so das Basissprachmodell einen ausreichenden Grad an Ausführlichkeit aufweist, übernommen werden kann.

20

Bestimmung der Wahrscheinlichkeit $P(w_i|C_i)$:

Die Wahrscheinlichkeit $P(w_i|C_i)$ für alle bezüglich des Basissprachmodells neuen Wörter w_i sind neu zu berechnen und die Wahrscheinlichkeit $P(w_i|C_i)$ (Wortwahrscheinlichkeit) des im Basissprachmodell vorhandenen Vokabulars ist vorzugsweise entsprechend anzupassen. Zur Bestimmung der Wahrscheinlichkeit $P(w_i|C_i)$ werden nachfolgend drei unterschiedliche Methoden erläutert:

30

Methode 1:

Die Wahrscheinlichkeit $P(w_i|C_i)$ für alle neuen Wörter w_i der neuen Domäne wird auf Basis des Textes für die neue Domäne abgeschätzt. Es wird ausgegangen von einem auf linguistischen

35

Klassen basierenden Basissprachmodell, wobei die neu abgeschätzte Wahrscheinlichkeit $P(w_i|C_i)$ vorzugsweise in das Basissprachmodell übernommen wird und somit eine Adaption des Basissprachmodells anhand des neuen Textes erfolgt.

- 5 Vorzugsweise wird diese Vorgehensweise verwendet, wenn der neue Text für die neue Domäne einen ausreichenden Umfang aufweist. Jedem Wort des neuen Textes wird die im Satzkontext ermittelte mindestens eine linguistische Klasse zugewiesen. Dies wird anhand des in [3] beschriebenen "Tagging-Tools"
- 10 durchgeführt. Die Wahrscheinlichkeit $P(w_i|C_i)$ für jedes neue Wort w_i wird anhand des neuen Textes abgeschätzt.

- In Fig.2 ist ein erstes Adaptionsverfahren zur Bestimmung der Wortwahrscheinlichkeit $P(w_i|C_i)$ dargestellt. Mittels des
- 15 Klassifikators F 311 und des linguistischen Lexikons 206 wird unter Verwendung des Tagging-Tools 202 (siehe detaillierte Erläuterungen zu dem Tagging-Tool unter [3]) sowohl aus einer Datenbasis aus großen Texten 201 ein großer "getaggtter" Text 203 als auch aus einer Datenbasis aus einem kleinen Text der
- 20 neuen Domäne (also dem neuen Text) 207 ein kleiner "getaggtter" Text 208 bestimmt. Aus dem großen "getaggtten" Text 203 wird mittels eines Sprachmodellgenerators 204 ein Basissprachmodell 205, das auf linguistischen Klassen beruht, bestimmt. Wie oben ausführlich beschrieben wurde, geht die
- 25 Wahrscheinlichkeit $P(C_i|C_{i-1})$ unverändert in das Sprachmodell für die neue Domäne ein. Aus dem "getaggtten" kleinen Text 208 wird mittels eines Adaptions-Tools 209, das eine Abschätzung der Wahrscheinlichkeit $P(w_i|C_i)$ mittels des "getaggtten" kleinen Textes durchführt, ein neues,
- 30 vorzugsweise adaptiertes, Sprachmodell 210 bestimmt. Neben der beschriebenen Adaption kann ohne Einschränkung auch ein weiteres Sprachmodell erzeugt werden.

Methode 2:

35

Ein weiteres Verfahren reduziert die durch das Basissprachmodell gegebenen einzelnen Wortwahrscheinlich-

keiten $P(w_i|C_i)$ und überträgt den reduzierten Beitrag auf das in der jeweiligen linguistischen Klasse C_i zu ergänzende Vokabular (Wörter). Dies wird abhängig von den jeweiligen Werten $P(w_i)$ der neuen Wörter durchgeführt.

5

In Fig.3 ist dieses zweite Adaptionsverfahren zur Bestimmung der Wahrscheinlichkeit $P(w_i|C_i)$ dargestellt. Mittels des Klassifikators F 311 und des linguistischen Lexikons 306 wird anhand des Tagging-Tools 302 aus einer Datenbasis aus großen Texten 301 ein "getaggtter" großer Text 303 bestimmt. Mit einem Sprachmodellgenerator 304 wird aus dem "getaggtten" großen Text 303 ein Basissprachmodell 305, das auf linguistischen Klassen beruht, erstellt. Aus dem Basissprachmodell 305 wird die Wahrscheinlichkeit $P(C_i|C_{i-1})$ unverändert übernommen. Anhand einer Wortliste für die neue Domäne 307 wird mittels eines Adaptions-Tools 308 ein adaptiertes Sprachmodell 309 bestimmt. Auch hier kann eine Adaption eine Veränderung oder eine Erzeugung eines Sprachmodells umfassen. Das Adaptions-Tool 308 berechnet die Wahrscheinlichkeit $P(w_i|C_i)$ für neue Wörter aus der Wahrscheinlichkeit $P(w_i)$ und führt eine Renormierung der Wahrscheinlichkeit $P(w_i|C_i)$ des Basissprachmodells durch. Nachfolgend wird dieses Verfahren detailliert beschrieben.

25 Die folgenden Größen sind durch das Basissprachmodell gegeben:

| | | |
|----|-------|---|
| 30 | w_i | Vokabular des Basissprachmodells mit $i=1, \dots, N,$ |
| | N | Anzahl des Vokabulars des Basissprachmodells, |
| 35 | N_C | Anzahl der Wörter in der jeweiligen Klasse C des Basissprachmodells, |
| | k | Anzahl der Klassen des Basissprachmodells, |

- $P(C_j)$ für $j=0, \dots, k$; Unigramm-
 Wahrscheinlichkeiten für die Klassen des
 Basissprachmodells,
- $P(C_j | C_{j-1})$ für $j=1, \dots, k$; Bigramm-Wahrscheinlichkeiten
 für die Klassen des Basissprachmodells,
- $P(w_i | C_j(w_i))$ und $P(C_j(w_i) | w_i)$
 Wortwahrscheinlichkeiten des
 Basissprachmodells
 für alle Wörter w_i , mit $i=1, \dots, N$,
 und alle Klassen $C_j(w_i)$ mit $j=0, \dots, k$,
 für die gilt: Wort w_i liegt in der
 linguistischen Klassen C_j .

Die Bildung linguistischer Klassen eines Sprachmodells für
 die neue Domäne entspricht der Bildung linguistischer Klassen
 für das Basissprachmodell. Der Klassifikator F (siehe
 Gleichung (2)) des Basissprachmodells wird übernommen. Also
 ist die Anzahl der linguistischen Klassen k unverändert. Es
 wird für die neue Domäne von Texten ähnlicher Struktur wie
 dem Basissprachmodell zugrundeliegende Trainingstexte
 ausgegangen. Die Wahrscheinlichkeit der Klassen-Bigramme
 $P(C_j | C_{j-1})$ und die Wahrscheinlichkeit von Klassen-Unigrammen
 $P(C_j)$ des Basissprachmodells bleiben unverändert.

Zu betrachten bleiben die Wahrscheinlichkeit $P(w_i | C_j(w_i))$ und
 die Wahrscheinlichkeit $P(C_j(w_i) | w_i)$, die sich jeweils auf
 einzelne Wörter beziehen. Die Wahrscheinlichkeit $P(w_i | C_j(w_i))$
 und die Wahrscheinlichkeit $P(C_j(w_i) | w_i)$ werden für die nicht
 im Basissprachmodell enthaltenen Wörter der neuen Domäne neu
 berechnet. Bereits vorhandene Wahrscheinlichkeiten für die
 Wörter des Basissprachmodells sind zu reskalieren.

Für die neue Domäne sind folgende Werte gegeben:

| | | |
|----|----------|---|
| 5 | w_h | Vokabular der neuen Domäne mit $h=1, \dots, L$, das nicht im Basissprachmodell enthalten ist. |
| 10 | L | Anzahl des Vokabulars im Sprachmodell für die neue Domäne (Zielsprachmodell), |
| | L_C | Anzahl der neuen Wörter w_h und der Wörter w_i des Basissprachmodells in der jeweiligen (linguistischen) Klasse C_j , |
| 15 | $P(w_h)$ | klassenunabhängige Wortwahrscheinlichkeiten in der neuen Domäne. |

Die Wahrscheinlichkeit $P(w_h)$ ist gegeben durch eine Wortliste
mit Worthäufigkeiten und die Größe des zugrundeliegenden
20 Textes.

Abschätzung der Wahrscheinlichkeit $P(C_j(w_h) | w_h)$:

Zur Abschätzung der Wahrscheinlichkeit $P(C_j(w_h) | w_h)$, jeweils
für das neue Vokabular w_h , wird davon ausgegangen, daß
25 $P(C_j(w_h) | w_h)$ annähernd gleich ist für Wörter, die in
derselben Klasse C_j liegen. Demnach gilt folgende Näherung:

$$P(C_j | w_h) \approx \frac{1}{N_C} \times \sum_{i=1}^{N_C} P(C_j | w_i) \quad (6),$$

30 wobei w_i alle Wörter des Basissprachmodells bezeichnet, die
in der Klasse C_j liegen.

Für die folgenden Ausführungen wird exemplarisch eine Klasse C_j untersucht. Zur vereinfachten Darstellung wird im folgenden diese Klasse C_j als Klasse C bezeichnet.

- 5 Vorhandene Wahrscheinlichkeitswerte sind zu renormieren.
Nachfolgend werden die "alten" Werte mit einer Tilde gekennzeichnet.

- 10 Eine weitere Näherungslösung für die Gleichung (6) ist die
Summe über die Wörter w_i , für die alle linguistischen Klassen mit den Klassen des neuen Wortes übereinstimmen.

Zur Berechnung der neuen Wahrscheinlichkeit $P(w_h | C(w_h))$ bzw. zur Renormierung der gegebenen Wahrscheinlichkeiten

- 15 $P(w_i | C(w_i))$ wird folgendes Verfahren angewandt:

- 1) Bestimmung des Anteils α der neuen Wörter im Wortschatz

- 20 2) Renormierung von

$$\tilde{P}(w) \quad (7)$$

mit:

25
$$P(w_i) = (1 - \alpha) \times \tilde{P}(w_i) \quad (8)$$

Bestimmung des Anteils α gemäß Gleichung (11)

- 30 3) Bestimmung des Anteils $1 - \gamma_C$ der neuen Wörter in der Klasse C ,

- 4) Bestimmung von $\tilde{P}(w_h | C)$ gemäß Gleichung (17),

- 35 5) Bestimmung von γ_C nach Gleichung (16),

6) Bestimmung der Wahrscheinlichkeit $P(w_h | C(w_h))$ durch:

$$P(w_h | C) = \frac{\gamma_C}{1 - \alpha} \tilde{P}(w_h | C) \quad (9),$$

5 siehe auch Gleichung (16). Der Faktor kann als Quotient der Anteile der alten Wörter in der linguistischen Klasse C und dem Anteil des alten Wortschatzes im Gesamtwortschatz interpretiert werden.

10 Zur Bestimmung des Anteils α folgt mit Gleichung (8):

$$\begin{aligned} 1 &= \sum_{i=1}^{N_C} \tilde{P}(w_i) = \sum_{i=1}^{N_C} P(w_i) + \sum_{h=N_C+1}^{L_C} P(w_h) = \\ &= (1 - \alpha) \times \sum_{i=1}^{N_C} \tilde{P}(w_i) + \sum_{h=N_C+1}^{L_C} P(w_h) \end{aligned} \quad (10),$$

daraus folgt:

$$\alpha = \sum_{h=N_C+1}^{L_C} P(w_h) \quad (11).$$

Zur Bestimmung des Anteils γ_C wird der Satz von Bayes auf
20 Gleichung (6) (siehe hierzu [5]) angewandt. Es folgt:

$$P(w_h | C) = \frac{P(w_h)}{N_C} \times \sum_{i=1}^{N_C} \frac{P(w_i | C)}{P(w_i)} \quad (12).$$

Anhand von Gleichung (12) und unter Anwendung der
25 Normierungseigenschaft für Wahrscheinlichkeitswerte (siehe [6]) kann gezeigt werden:

$$\sum_{i=1}^{N_C} P(w_i | C) + \sum_{h=N_C+1}^{L_C} \frac{P(w_h)}{N_C} \times \sum_{i=1}^{N_C} \frac{P(w_i | C)}{P(w_i)} = 1 \quad (13).$$

Für

5

$$P(w_i | C) = \gamma_C \times \tilde{P}(w_i | C) \quad (14)$$

mit $\gamma_C \leq 1$ und $\tilde{P}(w_i | C)$ als alte (nicht renormierte) Verteilung, gilt:

10

$$\left(1 + \sum_{h=N_C+1}^{L_C} \frac{P(w_h)}{N_C} \times \sum_{i=1}^{N_C} \frac{\tilde{P}(w_i | C)}{P(w_i)} \right) \times \gamma_C = 1 \quad (15).$$

Daraus folgt mit Gleichung (8):

15

$$\gamma_C = \left(1 + \frac{1}{1 - \alpha} \times \sum_{h=N_C+1}^{L_C} \tilde{P}(w_h | C) \right)^{-1} \quad (16)$$

mit

$$\tilde{P}(w_h | C) = \frac{P(w_h)}{N_C} \times \sum_{i=1}^{N_C} \frac{\tilde{P}(w_i | C)}{\tilde{P}(w_i)} \quad (17).$$

20

Methode 3:

Die Wahrscheinlichkeit $P(w_i | C_i)$ für bezüglich des
 25 Basissprachmodells neue Wörter w_i wird unter Verwendung einer
 entsprechenden Wortliste angenähert. Bei dieser
 Vorgehensweise wird die Lösungsstrategie von dem unter
 'Methode 2' beschriebenen Verfahren übernommen. Die hier
 nicht vorhandene Wahrscheinlichkeit $P(w_h)$ für die neuen

Wörter w_h wird angenähert. Dies erfolgt in Abhängigkeit zu einer Hauptkategorie HC des jeweiligen Wortes w_h . Es ergibt sich näherungsweise:

$$P(w_h) \approx \frac{1}{N_{HC}} \times \left(\sum_{w_i \text{ aus HC}} \left(\sum_{C_j(w_i)} P(w_i | C_j) \times P(C_j) \right) \right) \quad (18).$$

Hierzu werden Eigenschaften von Wörtern w_i des Basissprachmodells verwendet. N_{HC} ist eine Anzahl des Vokabulars des Basissprachmodells, das in der Hauptkategorie HC liegt. Die Summierung läuft über alle Klassen C_j des Basissprachmodells, zu denen das jeweilige Wort w_i gehört.

Bestimmung der Wahrscheinlichkeit $P(C_{i-1} | w_{i-1})$:

Schließlich wird die Wahrscheinlichkeit $P(C_{i-1} | w_{i-1})$ wie nachfolgend beschrieben bestimmt. Es sei angemerkt, daß in Gleichung (3) die Wahrscheinlichkeit $P(C_{i-1} | w_{i-1})$ einen Index 'i-1' aufweist, der nachfolgend zur Vereinfachung als Index i gesetzt wird.

Die Wahrscheinlichkeit $P(C_i | w_i)$ ergibt sich jeweils aus der Wahrscheinlichkeit $P(w_i | C_i)$, die wie oben beschrieben, bestimmt worden ist:

$$P(C_i | w_i) = K \times P(w_i | C_i) \times P(C_i) \quad (4)$$

mit einem Normalisierungsfaktor

$$K = \left(\sum_{C_i} P(w_i | C_i) \times P(C_i) \right)^{-1} \quad (5).$$

Die linguistische Klasse C_i durchläuft dabei alle für das Wort w_i möglichen linguistischen Klassen. Die Wahrscheinlichkeit $P(C_i)$ wird aus dem Basissprachmodell übernommen (Unigramm-Wahrscheinlichkeit für jeweilige

5 linguistische Klasse des Basissprachmodells).

Im Rahmen dieses Dokuments wurden folgende Veröffentlichungen zitiert:

- [1] G. Ruske: "Automatische Spracherkennung - Methoden der Klassifikation und Merkmalsextraktion", Oldenbourg
5 Verlag, München 1988, ISBN 3-486-20877-2, Seiten 1-10.
- [2] L. Rabiner, B.-H. Juang: "Fundamentals of Speech Recognition", Prentice Hall 1993, S.447-450.
- [3] P. Witschel: "Constructing Linguistic Oriented Language
10 Models for Large Vocabulary Speech Recognition", 3rd
EUROSPEECH 1993, Seiten 1199-1202.
- [4] F. Guethner, P. Maier: "Das CISLEX-Wörterbuchsystem",
CIS-Bericht 94-76-CIS, Universität München, 1994.
- [5] W. Feller: "An Introduction to Probability Theory and its Applications", John Wiley & Sons, 1976, Seiten 124,125.
- 15 [6] W. Feller: "An Introduction to Probability Theory and its Applications", John Wiley & Sons, 1976, Seite 22.

Patentansprüche

1. Verfahren zur Bestimmung einer Wahrscheinlichkeit für das Auftreten einer Folge von mindestens zwei Wörtern bei einer Spracherkennung durch einen Rechner,
- 5 a) bei dem eine Sprache linguistische Klassen aufweist:

$$(C_1, \dots, C_k) = F \left((f_1, v_{11}, \dots, v_{1j}) \dots (f_m, v_{m1}, \dots, v_{mj}) \right),$$

- 10 wobei
- f_m ein linguistisches Merkmal,
- m die Anzahl linguistischer Merkmale,
- $v_{m1} \dots v_{mj}$ die linguistischen Werte des linguistischen Merkmals f_m ,
- 15 j die Anzahl der linguistischen Werte,
- C_1, \dots, C_k die linguistische Klassen,
- k die Anzahl der linguistischen Klassen,
- F eine Abbildungsvorschrift (Klassifikator) von linguistischen Merkmalen und linguistischen Werten auf linguistische Klassen
- 20 bezeichnen
- b) bei dem einem Wort mindestens eine der linguistischen Klassen zugeordnet wird;
- 25 c) bei dem die Wahrscheinlichkeit für das Auftreten der Folge von mindestens zwei Wörtern bestimmt ist durch:

$$P(W) \approx \prod_{i=1}^n \sum_{C_i} \sum_{C_{i-1}} P(w_i | C_i) \times P(C_i | C_{i-1}) \times P(C_{i-1} | w_{i-1})$$

- 30 wobei
- $P(W)$ Wahrscheinlichkeit für das Auftreten der Folge von mindestens zwei Wörtern
- W die Folge von mindestens zwei Wörtern,
- w_i das i -te Wort der Folge W mit ($i=1..n$),

25

- n die Anzahl der Wörter w_i der Folge W ,
 C_i eine linguistische Klasse C , die zu einem Wort w_i gehört,
 C_{i-1} eine linguistische Klasse, die zu einem Wort w_{i-1} gehört,
5 $\sum C_i$ die Summe über alle linguistischen Klassen C , die zu einem Wort w_i gehören,
 $P(w_i|C_i)$ die bedingte Wortwahrscheinlichkeit,
 $P(C_i|C_{i-1})$ die Wahrscheinlichkeit für Bigramme
10 (auch: Klassen-Bigramme, Bigramm-Wahrscheinlichkeit),
 $P(C_{i-1}|w_{i-1})$ die bedingte Klassenwahrscheinlichkeit bezeichnen.
- 15 2. Verfahren nach Anspruch 1,
bei dem für einen Text anhand eines vorbestimmten Basissprachmodells die Wahrscheinlichkeit $P(C_i|C_{i-1})$ für den Text bestimmt wird, indem die Wahrscheinlichkeit $P(C_i|C_{i-1})$ aus dem Basissprachmodell übernommen wird.
20
3. Verfahren nach Anspruch 2,
bei dem die Wahrscheinlichkeit $P(w_i|C_i)$ nach mindestens einer der folgenden Möglichkeiten bestimmt wird:
a) die Wahrscheinlichkeit $P(w_i|C_i)$ wird anhand des Textes
25 bestimmt;
b) die Wahrscheinlichkeit $P(w_i|C_i)$ wird für ein Wort w_i mit Hilfe einer vorgegebenen Wahrscheinlichkeit $P(w_i)$ bestimmt;
c) die Wahrscheinlichkeit $P(w_i|C_i)$ wird unter Verwendung
30 einer Wortliste bestimmt.
4. Verfahren nach Anspruch 3,
bei dem anhand der ermittelten Wahrscheinlichkeit $P(w_i|C_i)$ das Basissprachmodell angepaßt wird.

5. Verfahren nach Anspruch 3 oder 4,
bei dem die Wahrscheinlichkeit $P(C_i|w_i)$ anhand der
Wahrscheinlichkeit $P(w_i|C_i)$ wie folgt bestimmt wird:

5

$$P(C_i|w_i) = K \times P(w_i|C_i) \times P(C_i),$$

wobei

$$K = \left(\sum_{C_i} P(w_i|C_i) \times P(C_i) \right)^{-1}$$

10

einen Normalisierungsfaktor bezeichnet.

6. Verfahren nach einem der vorhergehenden Ansprüche,
bei dem, wenn die Wahrscheinlichkeit $P(W)$ für das
Auftreten einer Folge von mindestens einem Wort oberhalb
einer vorgegebenen Schranke liegt, die entsprechende
Folge von mindestens einem Wort erkannt wird, ansonsten
eine vorgegebene Aktion durchgeführt wird.

15

7. Verfahren nach Anspruch 6,
bei dem die vorgegebene Aktion eine Ausgabe einer
Fehlermeldung oder Anlaß für einen Abbruch des Verfahrens
ist.

20

8. Verfahren nach einem der Ansprüche 2 bis 7,
bei dem sich der Text auf einen vorgegebenen
Anwendungsbereich bezieht.

25

1/2

FIG 1

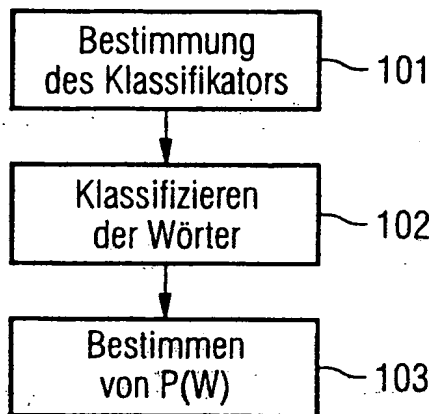
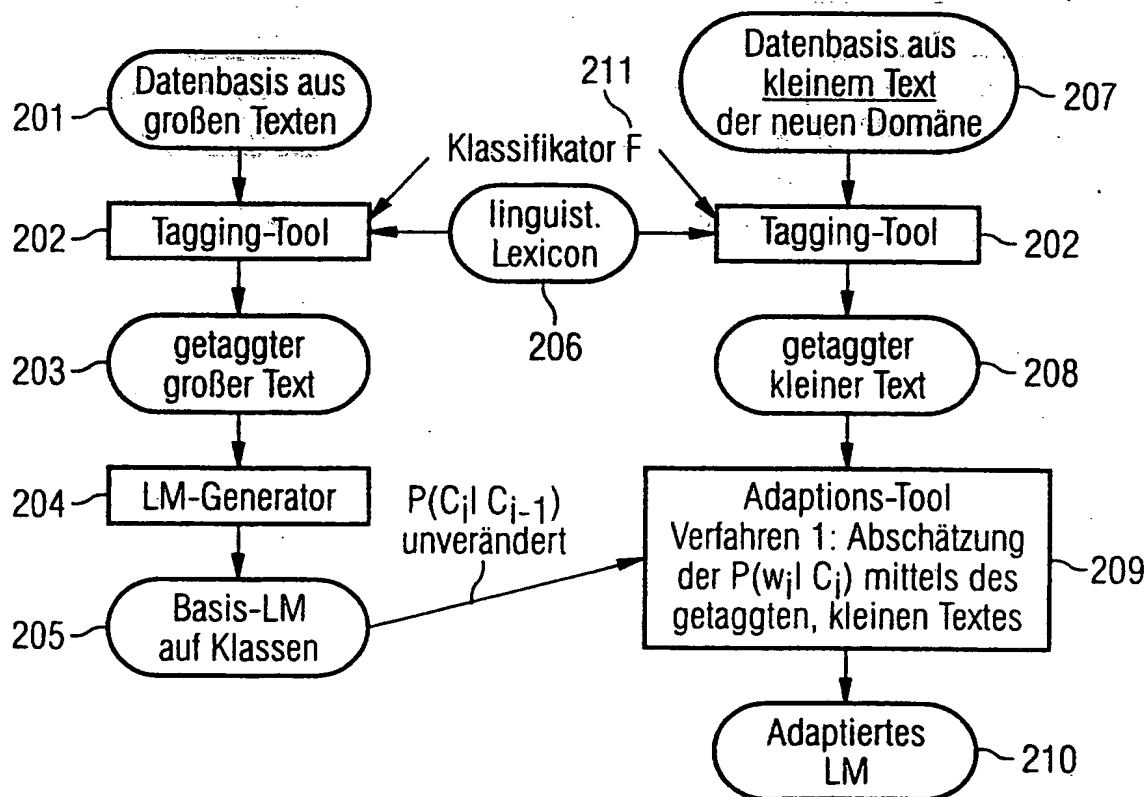
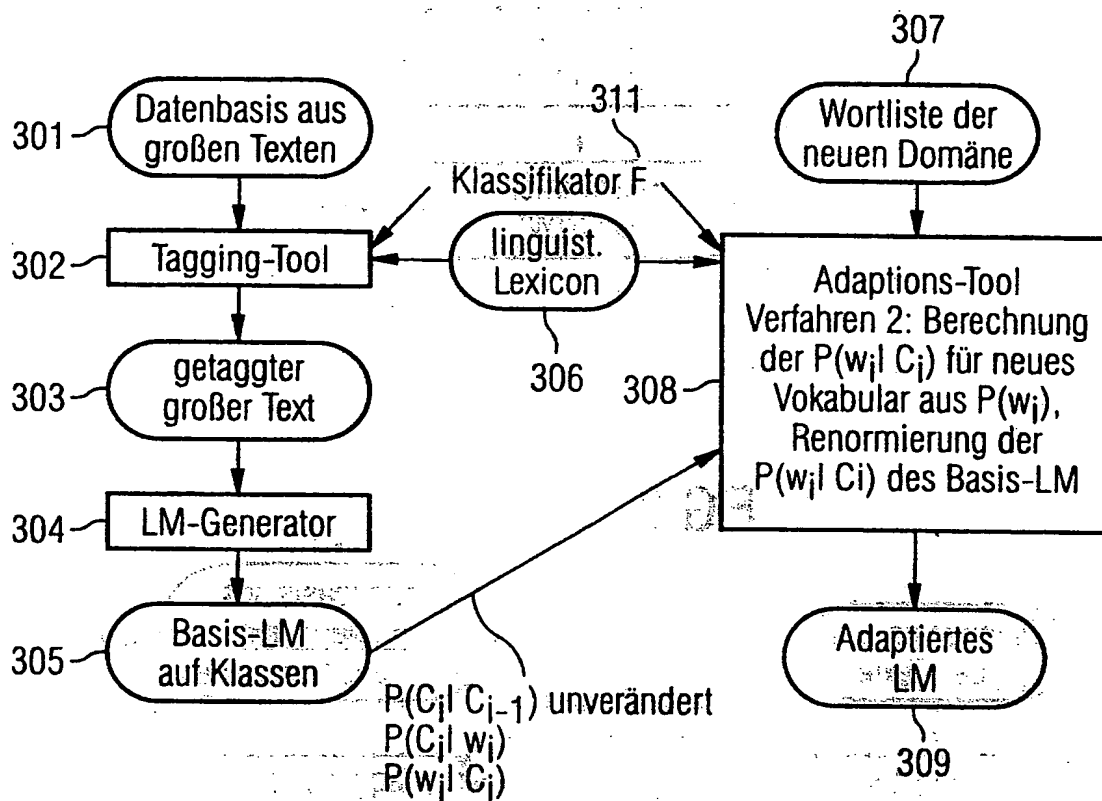


FIG 2



2/2

FIG 3



INTERNATIONAL SEARCH REPORT

In. ational Application No

PCT/DE 98/02632

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G10L5/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC 6 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category * | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|------------|--|-----------------------|
| A | <p>ELBEZE M ET AL: "A MORPHOLOGICAL MODEL FOR LARGE VOCABULARY SPEECH RECOGNITION" SPEECH PROCESSING 1, ALBUQUERQUE, APRIL 3 - 6, 1990, vol. 1, no. CONF. 15, 3 April 1990, pages 577-580, XP000146534 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS see page 578, left-hand column - right-hand column, line 3</p> <p style="text-align: center;">---</p> <p style="text-align: center;">-/--</p> | 1 |

☒ Further documents are listed in the continuation of box C.

☐ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

15 February 1999

Date of mailing of the international search report

22/02/1999

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.
Fax: (+31-70) 340-3016

Authorized officer

Pulluard, R

INTERNATIONAL SEARCH REPORT

International Application No

PCT/DE 98/02632

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

| Category * | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|------------|---|-----------------------|
| A | <p>MALTESE G ET AL: "AN AUTOMATIC TECHNIQUE TO INCLUDE GRAMMATICAL AND MORPHOLOGICAL INFORMATION IN A TRIGRAM-BASED STATISTICAL LANGUAGE MODEL"</p> <p>SPEECH PROCESSING 1, SAN FRANCISCO, MAR. 23 - 26, 1992,</p> <p>vol. 1, no. CONF. 17, 23 March 1992, pages 157-160, XP000341107</p> <p>INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS</p> <p>see page 157, right-hand column</p> | 1 |
| A | <p>JARDINO M ET AL: "AUTOMATIC WORD CLASSIFICATION USING SIMULATED ANNEALING"</p> <p>SPEECH PROCESSING, MINNEAPOLIS, APR. 27 - 30, 1993,</p> <p>vol. 2, 27 April 1993, pages II-41-44, XP000427720</p> <p>INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS</p> <p>see page 41, right-hand column, line 20 - page 42, left-hand column, line 53</p> | 1 |

INTERNATIONALER RECHERCHENBERICHT

In ationales Aktenzeichen

PCT/DE 98/02632

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 6 G10L5/06

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)

IPK 6 G10L

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

| Kategorie | Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile | Betr. Anspruch Nr. |
|-----------|--|--------------------|
| A | ELBEZE M ET AL: "A MORPHOLOGICAL MODEL FOR LARGE VOCABULARY SPEECH RECOGNITION" SPEECH PROCESSING 1, ALBUQUERQUE, APRIL 3 - 6, 1990, Bd. 1, Nr. CONF. 15, 3. April 1990, Seiten 577-580, XP000146534 INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS siehe Seite 578, linke Spalte - rechte Spalte, Zeile 3 --- -/-- | 1 |

☒ Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

☐ Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :

"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist

"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist

"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)

"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht

"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist

"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist

"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden

"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist

"Z" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche

15. Februar 1999

Absenddatum des internationalen Recherchenberichts

22/02/1999

Name und Postanschrift der Internationalen Recherchenbehörde
Europäisches Patentamt, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Bevollmächtigter Bediensteter

Pulluard, R

INTERNATIONALER RECHERCHENBERICHT

II. Internationales Aktenzeichen

PCT/DE 98/02632

C.(Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

| Kategorie | Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile | Beitr. Anspruch Nr. |
|-----------|--|---------------------|
| A | <p>MALTESE G ET AL: "AN AUTOMATIC TECHNIQUE TO INCLUDE GRAMMATICAL AND MORPHOLOGICAL INFORMATION IN A TRIGRAM-BASED STATISTICAL LANGUAGE MODEL"</p> <p>SPEECH PROCESSING 1, SAN FRANCISCO, MAR. 23 - 26, 1992,</p> <p>Bd. 1, Nr. CONF. 17, 23. März 1992, Seiten 157-160, XP000341107</p> <p>INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS</p> <p>siehe Seite 157, rechte Spalte</p> | 1 |
| A | <p>JARDINO M ET AL: "AUTOMATIC WORD CLASSIFICATION USING SIMULATED ANNEALING" SPEECH PROCESSING, MINNEAPOLIS, APR. 27 - 30, 1993,</p> <p>Bd. 2, 27. April 1993, Seiten II-41-44, XP000427720</p> <p>INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS</p> <p>siehe Seite 41, rechte Spalte, Zeile 20 - Seite 42, linke Spalte, Zeile 53</p> | 1 |